Segmenting Precipitation Series

Robert Lund

Department of Mathematical Sciences
Clemson University
Clemson, SC 29634-0975
Lund@clemson.edu

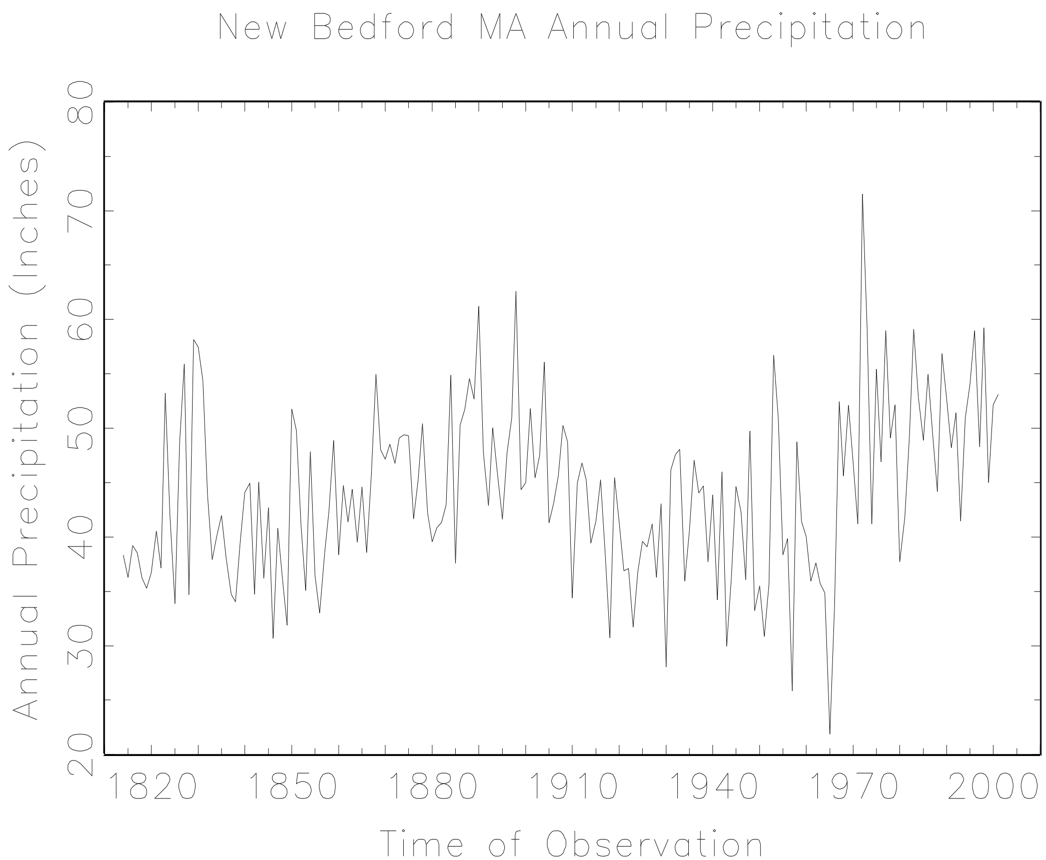
Why segment a precipitation series?

- Required before estimating trends.
- Many breakpoints are undocumented.
- Quality check data.
- Calibration of new gauges.

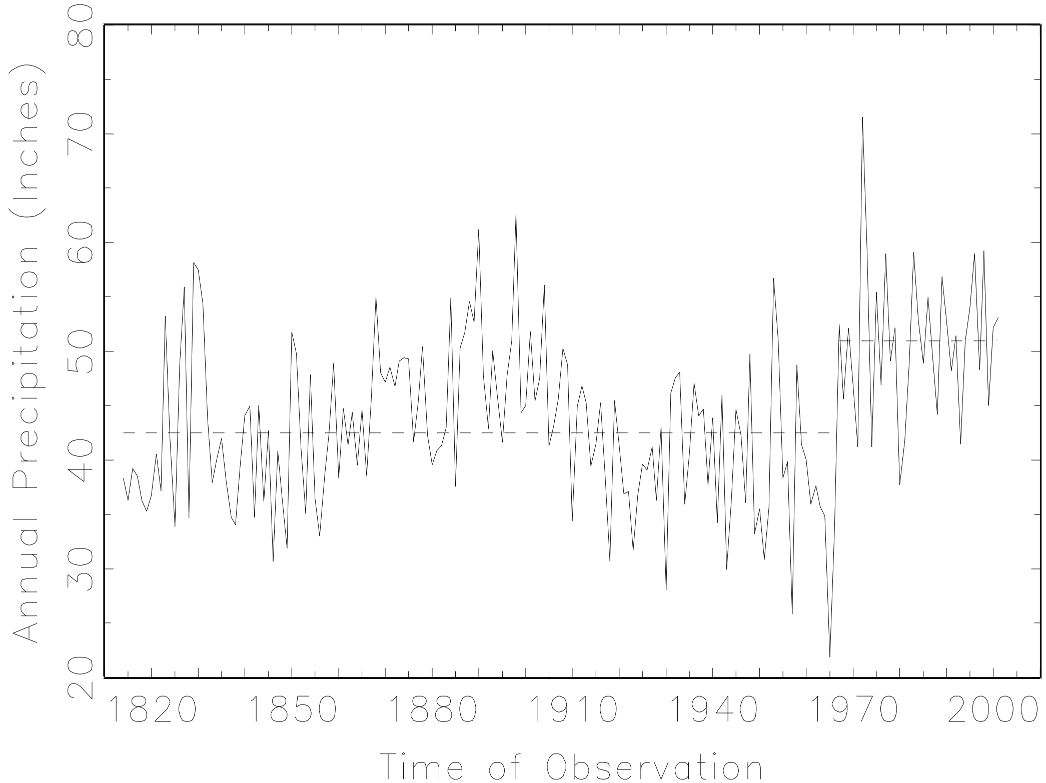
Key Questions

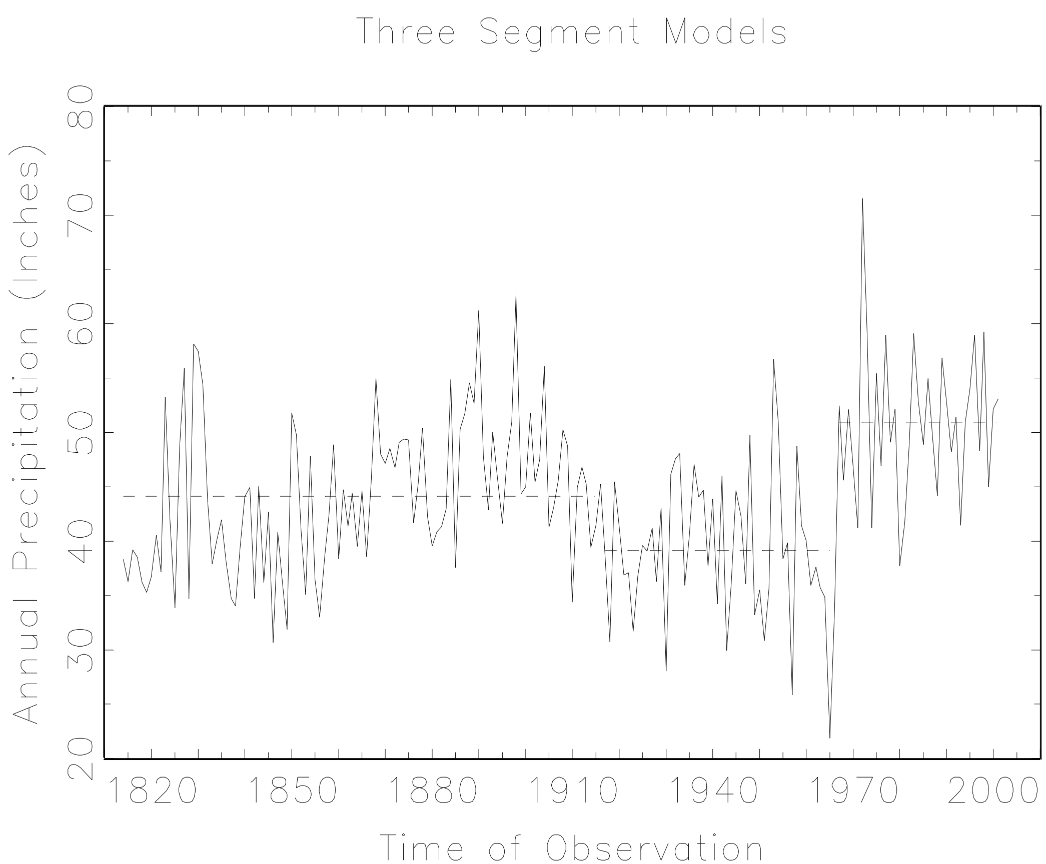
1. How many breakpoints are there?
2. Where are the breakpoints?

We examine 188 years of annual precipitations from New Bedford Massachusetts, USA, to illustrate the methods.


New Bedford MA Annual Precipitation
Annual Precipitation (Inches)
20
30
40
50
60
70
80
1820
1850
1880
1910
1940
1970
2000
Time of Observation

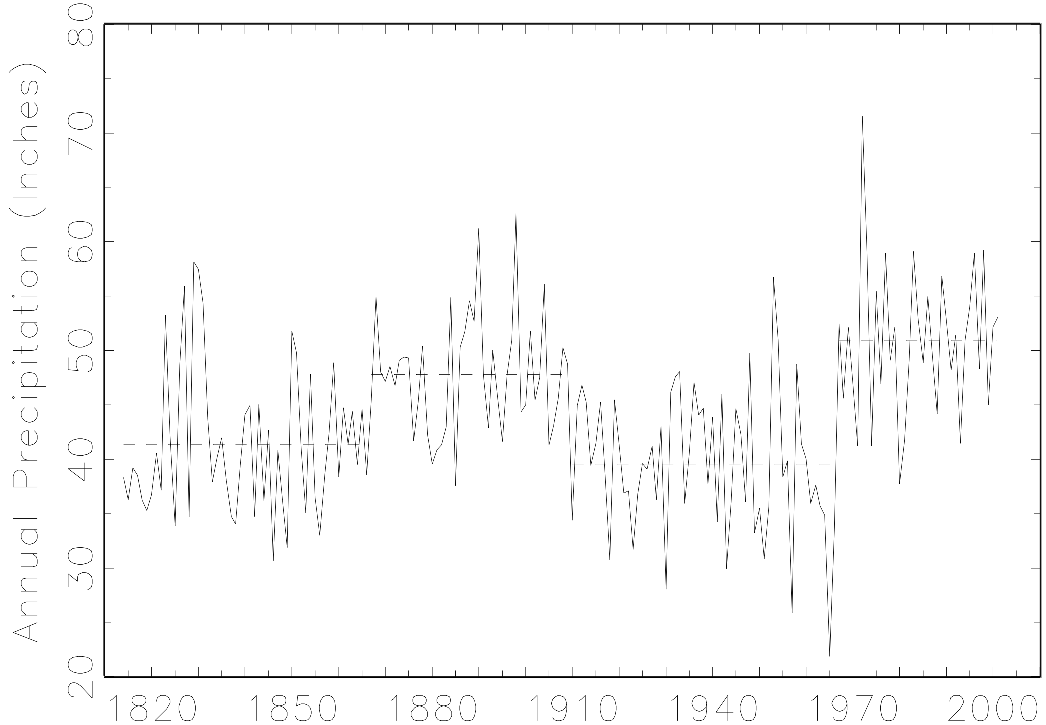
Two Segment Models

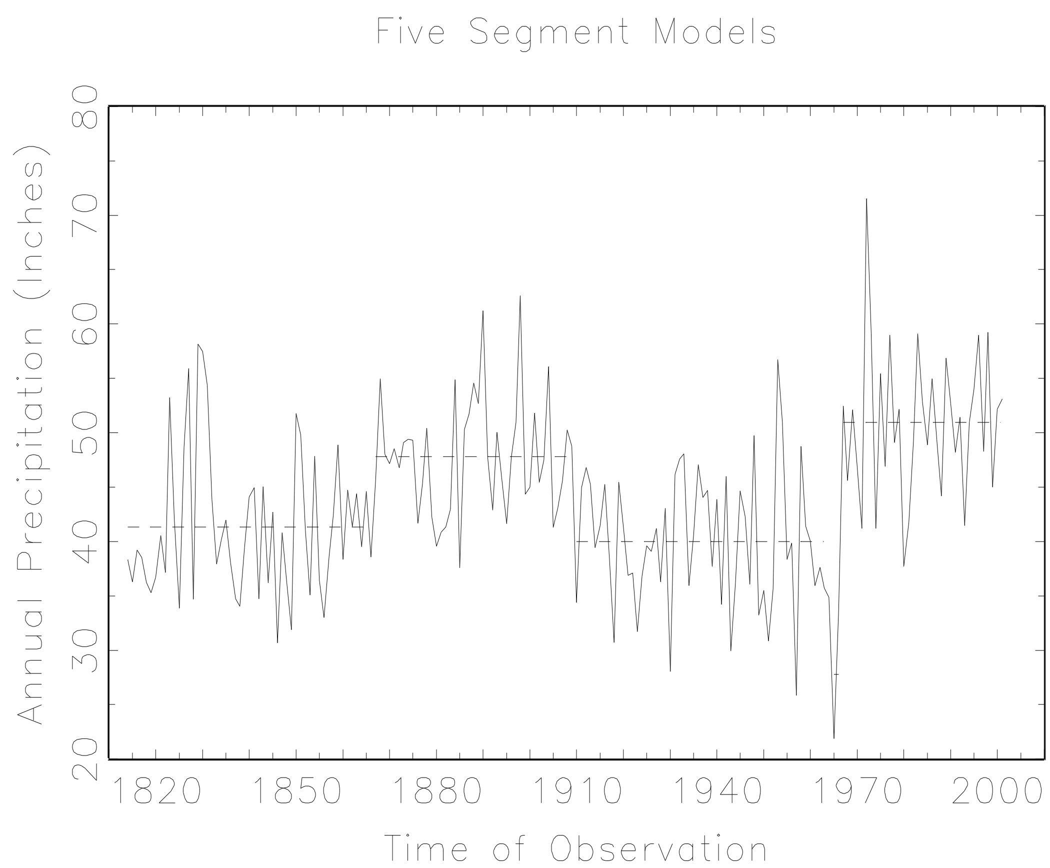


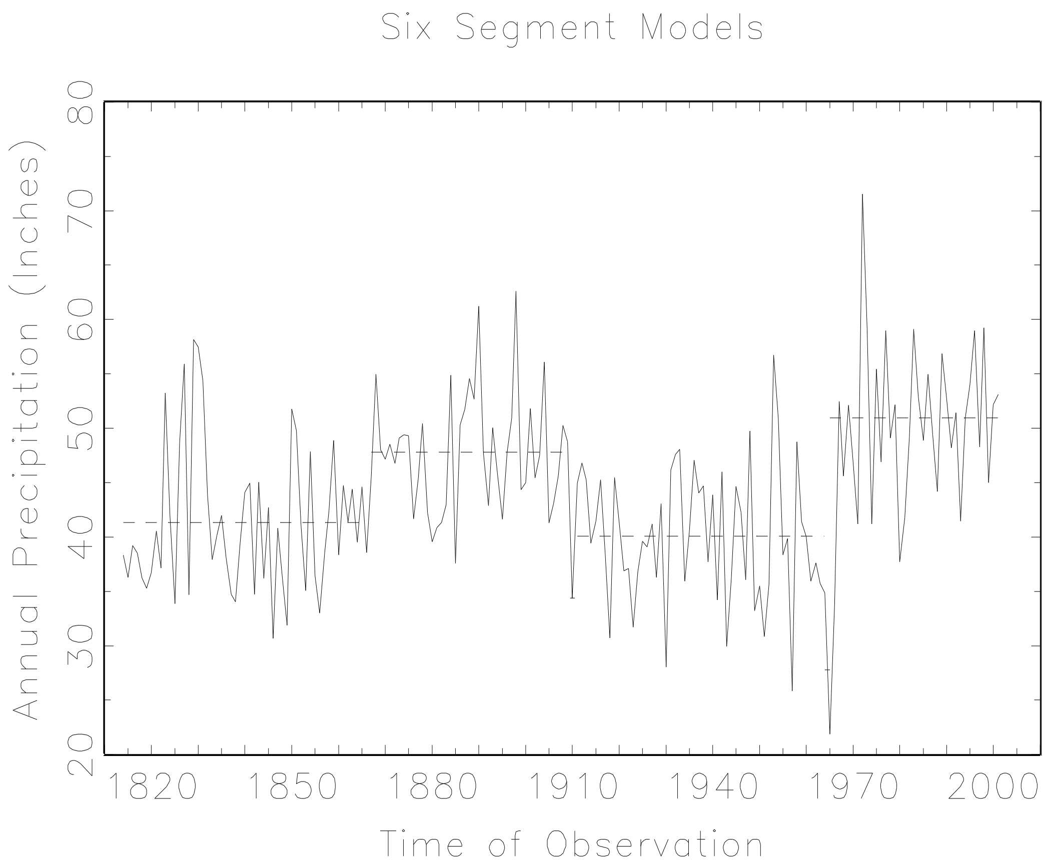


Three Segment Models
Annual Precipitation (Inches)
20
30
40
50
60
70
80
1820
1850
1880
1910
1940
1970
2000
Time of Observation

Four Segment Models




Five Segment Models
Annual Precipitation (Inches)
20
30
40
50
60
70
80
1820
1850
1880
1910
1940
1970
2000
Time of Observation



Six Segment Models
Annual Precipitation (Inches)
20
30
40
50
60
70
80
1820
1850
1880
1910
1940
1970
2000
Time of Observation

Model selection: MDL versus AIC

# Segments	Changepoint Times	AIC Score	MDL Score
2	154	371.1786	377.0333
3	104, 154	363.2085	375.6150
4	54, 97, 154	354.4078	373.0704
5	54, 97, 152, 154	349.9269	374.7318
6	54, 97, 98, 152, 154	351.4389	382.3217

Both AIC (Aikaike Information Criterion) and MDL (Minimum Description Length) select the model that minimizes the penalized likelihood score.

AIC prefers 5 segments, MDL 4.

AIC is notorious for overparametrizing.

The Mathematics behind the MDL Criterion

MDL criteria come from information theory.

The goal is to minimize a penalized likelihood.

MDL penalizes $\log_2(n)/2$ for each mean parameter and the overall variance, and $\log_2(n)$ for each integer-valued changepoint time and the number of changepoints.

The $\log_2(n)$ MDL penalty for the integer parameters is greater than that for AIC.

Details, details, details

Suppose the changepoint times and locations are known. The annual precipitations are modeled as lognormal at time t with mean $\mu_{R(t)}$, where $R(t)$ denotes the segment number $(1 \leq R(t) \leq m+1)$ of the data point at time t. The marginal density of X_t is

$$f(x_t) = \frac{1}{x_t \sigma \sqrt{2\pi}} \exp\left\{ -\frac{(\ln(x_t) - \mu_{R(t)})^2}{2\sigma^2} \right\}.$$

We assume the data from different years are independent.

Details, details, details

The likelihood function L of all n observations is

$$L = \prod_{t=1}^{n} f(X_t).$$

For known changepoint times (say m) and numbers, the parameter estimates are

$$\hat{\mu}_\ell = \frac{1}{\#(S_\ell)} \sum_{t \in S_\ell} \ln(X_t), \quad 1 \le \ell \le m+1;$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} (\ln(X_t) - \hat{\mu}_{R(t)})^2.$$

Here, S_ℓ is the set of all t where the series obeyed segment ℓ.

Details, details, details

MDL minimizes

$$\frac{N \ln(\hat{\sigma}^2)}{2} + \sum_{i=1}^{m+1} \#(S_i)\hat{\mu}_i + \frac{3m \ln(n)}{2} + \ln(m).$$

AIC minimizes

$$\frac{N \ln(\hat{\sigma}^2)}{2} + \sum_{i=1}^{m+1} \#(S_i)\hat{\mu}_i + 2m.$$

The MDL penalty is greater than the AIC penalty.

A true optimization requires that we search over all changepoint configurations and orders. This is hard to do but has recently become possible with the advance of genetic algorithms.